DOCUMENT RESUME

ED 381 556 TM 022 891

AUTHOR Livingston, Samuel A.; Sims-Gunzenhauser, Alice
TITLE Setting Standards on the Assessor Proficiency Test for the Praxis III: Classroom Performance Assessment.
INSTITUTION Educational Testing Service, Princeton, N.J.
REPORT NO ETS-RR-94-50
PUB DATE Nov 94
NOTE 23p.
PUB TYPE Reports - Research/Technical (143)

EDRS PRICE MF01/PC01 Plus Postage.
DESCRIPTORS *Beginning Teachers; *Criteria; Documentation; Educational Assessment; Elementary Secondary Education; *Evaluators; Licensing Examinations (Professions); Observation; Regression (Statistics); *Scoring; Standards; *Teacher Evaluation; Training
IDENTIFIERS Accuracy; Performance Based Evaluation; *Praxis III; Standard Setting

ABSTRACT

Praxis III is an assessment procedure that provides information for making instructional and licensing decisions about beginning teachers. The Praxis III Assessor's job is to interview the beginning teacher, observe the teacher in the classroom, score the teacher's performance on 19 criteria, and summarize the evidence for each score. The Assessor Proficiency Test (APT) consists of performing the assessor's task for a videotaped lesson. It yields two scores: accuracy and documentation. In this study, 5 judges (developers of the Praxis III assessment) evaluated the accuracy and documentation of APT records produced by 15 assessor trainees. Documentation judgments were made by group consensus; accuracy judgments were individual. Logistic regression analysis showed that both types of judgments were strongly related to the corresponding APT scores, but the judges' individual standards for accuracy varied greatly. Seven tables and two figures present study data. (Author/SLD)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

RR-94-50

ED 381 556

RESEARCH REPORT

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

☑ This document has been reproduced as received from the person or organization originating it.

☐ Minor changes have been made to improve reproduction quality.

• Points of view or opinions stated in this document do not necessarily represent official OERI position or policy.

"PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY

R. Coley

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)."

SETTING STANDARDS ON THE ASSESSOR PROFICIENCY TEST FOR THE PRAXIS III: CLASSROOM PERFORMANCE ASSESSMENT

Samuel A. Livingston
Alice Sims-Gunzenhauser

Educational Testing Service
Princeton, New Jersey
November 1994

TM022891

BEST COPY AVAILABLE

Setting Standards on the Assessor Proficiency Test for the Praxis III: Classroom Performance Assessment

Samuel A. Livingston
Alice Sims-Gunzenhauser

Educational Testing Service

Copyright © 1994. Educational Testing Service. All rights reserved.

Setting Standards on the Assessor Proficiency Test for the Praxis III: Classroom Performance Assessment

Samuel A. Livingston
Alice Sims-Gunzenhauser

Educational Testing Service

Abstract

The Praxis III Assessor's job is to interview the beginning teacher, observe the teacher in the classroom, score the teacher's performance on nineteen criteria, and summarize the evidence for each score. The Assessor Proficiency Test (APT) consists of performing the assessor's task for a videotaped lesson. It yields two scores: accuracy and documentation. In this study, five judges judged the accuracy and documentation of APT records produced by fifteen assessor trainees. Documentation judgments were made by group consensus; accuracy judgments were individual. Logistic regression analysis showed that both types of judgments were strongly related to the corresponding APT scores, but judges' individual standards for accuracy varied greatly.

Setting Standards on the Assessor Proficiency Test for the Praxis III: Classroom Performance Assessment

Samuel A. Livingston
Alice Sims-Gunzenhauser

Educational Testing Service

The Praxis III assessor

Praxis III is an assessment procedure that provides information for making instructional and licensing decisions about beginning teachers. Praxis III consists of a series of classroom observations, each preceded and followed by an interview with the beginning teacher. The person who conducts the interviews and observes the beginning teacher's performance in the classroom is called the assessor. After each observation and the accompanying interviews, the assessor completes a Record of Evidence form, assigning a separate score to the beginning teacher's performance on each of nineteen criteria. Each of the nineteen criteria focuses on a specific aspect of the beginning teacher's performance. The criterion scores are expressed on a scale with six levels: 1.0, 1.5, 2.0, 2.5, 3.0, and 3.5. The nineteen criteria are divided into four domains. The first three domains each include five criteria; the fourth domain includes only four. Table 1 lists the nineteen criteria.

During the interviews and the observations, the assessor takes notes on the beginning teacher's performance. In completing the Record of Evidence form, the assessor reviews these notes and decides what score best characterizes the beginning teacher's performance on each criterion. Each criterion score represents a judgment by the assessor, based on events that occurred in the classroom or in the interviews. The assessor must be able to support these judgments convincingly. Therefore, the Record of Evidence form contains spaces in which the assessor must provide two kinds of narrative information for each criterion: (1) specific evidence from the observation and interviews that serves as the basis for the assigned score, and (2) a summary statement of the teacher's performance on the criterion.

Clearly, the assessor's role in Praxis III is central. The validity of the Praxis III assessment and the fairness of the decisions based on it depend directly on the competence of the assessors.[1] Therefore, the qualifications and the training of the assessors are critically important for Praxis III. The individuals selected to become assessors are experienced local educators, selected by their state or other organization. To become assessors, they must

---

[1]Educational Testing Service recommends that, to the extent possible, each observation of the beginning teacher's performance be conducted by a different assessor. Any decision about a beginning teacher should be based on observations made on at least two different occasions and by at least two different assessors. (See Guidelines for Proper Use of The Praxis Series: Professional Assessments for Beginning Teachers™.)

complete a five-day training course, in which they practice interviewing, observing taped classroom events, taking notes, identifying the Praxis III criteria relevant to a classroom event, using the Praxis III scale, and writing evidence and summary statements. In addition, they complete a field experience in which they conduct a complete Praxis III assessment cycle, consisting of an observation, the accompanying interviews, and the completion of the Record of Evidence form.

The Assessor Proficiency Test

The assessor training concludes with the Assessor Proficiency Test. Its purpose is to make sure that every assessor who goes into a school to conduct a Praxis III assessment can correctly apply the scoring rules and complete the Record of Evidence form. The test is based on a videotape of a Praxis III assessment cycle, including the pre-observation interview, the teacher's performance in the classroom, and the post-observation interview. The assessor trainee watches the videotape and completes the Record of Evidence form. Scores on the Assessor Proficiency Test are based on the contents of the completed Record of Evidence form.

Each assessor trainee receives two scores on the Assessor Proficiency Test: an accuracy score and a documentation score. Both scores are expressed on a scale of 0 to 100. The accuracy score is a measure of the extent to which the assessor trainee has correctly applied the scoring rules. The documentation score is a measure of the extent to which the assessor trainee has provided satisfactory evidence and an adequate summary statement for each criterion. Both scores are necessary to determine whether the trainee can perform acceptably as an assessor.

The accuracy score is computed from the criterion scores the assessor trainee assigns to the videotaped performance. For each criterion in the videotaped performance there is a juried criterion score that serves as the correct answer. The juried criterion scores were determined by a consensus process involving the developers of Praxis III (including ETS staff and practicing teachers). The accuracy score is based on the differences between the criterion scores assigned by the assessor trainee and the juried criterion scores.

The documentation score is based on the written evidence and summary statements the assessor trainee writes on the Record of Evidence form. For each of the nineteen criteria, the rater of the Assessor Proficiency Test assigns a rating of 1, 2, 3, or 4 points for the written evidence and a rating of 1, 2, 3, or 4 points for the summary statement. The rater also classifies the assessor's evidence and summary as objective (2 points) or not objective (no points). The documentation score is based on these ratings.

The purpose of the study

At the time this study was conducted, the Praxis III developers had already determined that the accuracy score would be a function of the differences between the assessor's assigned criterion scores and the juried criterion scores. They had not yet determined the formula by which these

differences would be combined into a single numerical score. The present study was intended to help the Praxis III developers choose a scoring formula and a point on the resulting score scale so as to discriminate between Record of Evidence forms on which the accuracy of the criterion scores was acceptable and forms on which it was not.

Similarly, at the time this study was conducted, the Praxis III developers had already determined that the documentation score would be a function of the ratings for evidence, summary statement, and objectivity on the nineteen criteria. They had not yet determined the formula by which these ratings would be combined into a single numerical score. The study was intended to provide information that would enable the Praxis III developers to choose a scoring formula and a point on the resulting score scale so as to discriminate between Record of Evidence forms on which the documentation represented acceptable performance by the assessor and forms on which it did not.

The procedure

For both the accuracy and documentation scores, the basis for the selection of a scoring formula and for the choice of a qualifying score was the holistic judgment of performance by actual assessor trainees on the Assessor Proficiency Test. Performance, in this context, means the accuracy of the criterion scores assigned by the assessor trainees and the adequacy of their written documentation. The study compared different methods of computing accuracy and documentation scores in terms of their agreement with the holistic judgments.[2] For the scoring methods selected, the study produced estimates of the probability of a favorable holistic judgment, as a function of the score.

The judges for this study were five developers of the Praxis III assessment. Their roles in the development process differed. Judge 1 was the project director. Judge 2 was the developer and co-ordinator of the assessor training course. Judge 3 was an assessor training leader who also served as a supervisor and mentor to assessors. Judge 4 was the developer of the forms and procedures used to document the assessment process. Judge 5 was an assessor training leader and the developer of videotapes used in training assessors. All five judges were thoroughly familiar with the videotaped performance used in the test.

The procedure for the study consisted of the following steps:

1. One of the Praxis III developers selected a sample of fifteen Record of Evidence forms completed by assessor trainees taking the Assessor Proficiency Test. The Record of Evidence forms were selected to represent a wide range of quality and a varied selection of problems observed from assessor trainees taking the test.

---

[2]This portion of the study could be described as a "policy capturing" study.

2. This same Praxis III developer rated the documentation on each of the fifteen Record of Evidence forms, assigning ratings for evidence, summary, and objectivity on each of the nineteen criteria.

3. The five judges individually reviewed the fifteen Record of Evidence forms and made two holistic yes-or-no judgments. The first was an accuracy judgment: whether the criterion scores on the Record of Evidence form were acceptably close to the juried criterion scores. The second was a documentation judgment: whether the evidence and summary statements on the form reflected acceptable performance as an assessor. Each of the judges was given the Record of Evidence forms in a different, randomly determined sequence. The judges were not given any of the information resulting from the rating of the documentation on the Record of Evidence forms.

4. Four of the five judges met to discuss their judgments of the documentation, resolve disagreements, and reach a group consensus judgment of the documentation on each Record of Evidence form. The Praxis III developer who selected and scored the Record of Evidence forms and the statistician who would analyze the resulting data were present at this meeting but did not participate in the process of resolving disagreements about specific Record of Evidence forms.

5. After the group discussion, the judges reviewed the consensus judgments and noted any Record of Evidence forms on which they disagreed with the group consensus judgment.

Because of time limitations, it was not possible to get consensus judgments for both accuracy and documentation. The designers of the study gave priority to consensus judgments for documentation because much of the judgment required to evaluate accuracy had been incorporated into the process of determining the juried criterion scores.

The judgments

The upper portion of Table 2 shows which judges classified each Record of Evidence form as acceptably accurate. In this table the individual judgments are represented by 1 (acceptable) or 0 (unacceptable). "Number OK" is the number of judgments of a Record of Evidence form as acceptable. The Record of Evidence forms are listed in order of the number of favorable judgments they received (low to high). The judges are listed in order of the number of favorable judgments they awarded (high to low). The "Record ID" and "Judge ID" numbers are arbitrary; they are included to provide a link between Tables 2 and 3. The accuracy standards of the individual judges varied greatly. Judge 3 classified thirteen of the fifteen Record of Evidence forms as acceptable for accuracy; Judge 4 classified only four of the fifteen as acceptable. The lower portion of Table 2 shows the correlations[3] between the accuracy judgments made by the individual judges. These correlations range from a low of -.04 to a high of .66, indicating only moderate agreement among

[3]All correlations in this report are product-moment correlations.

the individual judges as to which particular Record of Evidence forms were acceptably accurate.

The upper portion of Table 3 shows which judges classified the documentation on each Record of Evidence form as acceptable, before the group discussion; it also shows the group consensus judgment. The Record of Evidence forms and the judges are listed in order of the number of favorable judgments, as in Table 2. For twelve of the fifteen Record of Evidence forms, the group consensus judgment for documentation agreed with the majority of the individual judgments made before discussion. Two of the three exceptions were Record of Evidence forms on which the judges were originally divided three-to-two. However, there was one Record of Evidence form that was judged as acceptable by only one judge before the discussion but was judged acceptable by the group after discussion. There were three Record of Evidence forms for which one judge dissented from the group judgment.

The lower portion of Table 3 shows the correlations between the judgments by the individual judges and by the group. The correlations between the individual judges' judgments before discussion range from a low of -.17 to a high of only .38, indicating substantial disagreement among the individual judges. The correlations of the individual judgments before discussion with the group judgments after discussion range from .05 to .50.

There was a strong tendency for the same Record of Evidence forms to be judged acceptable for both accuracy and documentation ($r = .74$). That is, the Record of Evidence forms that most judges classified as acceptably accurate tended to be the same ones that they classified as containing adequate documentation. There was also a strong tendency for the same judges to be strict or lenient in judging both characteristics ($r = .77$). Judge 4 was strict in judging both accuracy and documentation; Judges 1 and 3 were lenient.

The scores

The computation of an accuracy score begins with the differences between the criterion scores assigned by the assessor trainee and the juried criterion scores. Each difference is assigned a number of penalty points, which in most cases is equal to the size of the difference, up to a maximum of 2.0. (A difference greater than 2.0 was impossible on most criteria, given the juried ratings for this videotape.) The number of penalty points is summed over the nineteen criteria, and the result is divided by the maximum possible value for that sum. This fraction is then subtracted from 1.00, and the result is multiplied by 100, to produce a score that can vary from 0 to 100, with zero representing the worst possible performance and 100 the best.

The Praxis III developers were undecided about two issues in computing the accuracy scores. One issue was whether or not to penalize a difference of 0.5 between the assessor trainee's assigned criterion score and the juried criterion score. The accuracy score that disregarded differences of 0.5 reflected the project director's opinion that competent assessors could reasonably award criterion scores that differed by 0.5. The project director's opinion was based on the absence of scale descriptors for criterion

scores of 1.5, 2.5, and 3.5 and also on the observed agreement between assessors in tryouts of early versions of the assessment. The analysis of the data from the present study included accuracy scores computed both ways: with and without a penalty for a difference of 0.5.

The second issue in computing the accuracy scores involved the weighting of the nineteen Praxis III criteria. A score that gives all nineteen criteria equal weight has the advantage of simplicity. However, some of the criteria may be difficult to apply to a particular videotaped performance. For the videotaped performance used in this study, the Praxis III developers identified three such criteria. Again, the analysis included accuracy scores computed both ways: with equally weighted criteria and with unequally weighted criteria. In the unequal weighting, the weights for the three hard-to-apply criteria were reduced by half.

The documentation score is computed from the ratings for evidence, summary statement, and objectivity on each criterion. One issue in computing the documentation scores was how to use the ratings to arrive at a score. One way is to make the score a function of the total number of rating points awarded, transforming this total onto a scale on which 0 represents the worst possible performance and 100 the best. A second way, suggested by one of the Praxis III developers, is to use the ratings to make a pass/fail decision for each criterion and to base the score on the number of criteria passed. The documentation for a criterion "passed" if the ratings for evidence, summary statement, and objectivity summed to at least seven, with ratings of at least two for both the evidence and the summary statement. The documentation score computed by this method was simply the percentage of criteria passed.

As with the accuracy scores, a second issue in computing the documentation scores was the weighting of the criteria. The analysis included documentation scores computed with equally weighted criteria and with unequally weighted criteria.

Comparing scores produced by different methods

Table 4 compares the four ways of computing accuracy scores, in terms of statistics describing the scores of the fifteen Record of Evidence forms included in the study. The upper portion of the table shows the highest and lowest scores and the mean and standard deviation of the scores computed by each procedure. Disregarding a difference of 0.5 from the juried criterion score had the predictable effect of producing higher accuracy scores, by about twelve or thirteen points on the average. The variation in the scores of the fifteen Record of Evidence forms was about the same for both methods. Weighting the criteria unequally tended to produce scores that varied somewhat less than the scores based on equally weighted criteria.

The lower portion of Table 4 shows the correlations of the accuracy scores with the judgments by each judge. Disregarding differences of 0.5 did not appear to have a systematic effect on the correlations between the scores and the judgments. Weighting the nineteen criteria unequally tended to produce slightly lower correlations between the scores and the judgments.

Table 5 compares the different methods of computing documentation scores. The method based on the number of criteria passed tended to produce a wider range of scores than the method based on the total number of rating points awarded. The scores based on the number of criteria passed were higher, on the average, and tended to vary much more. Weighting the criteria unequally had little effect.

The lower portion of Table 5 shows the correlations of the documentation scores with the judgments. Regardless of the method of computing the scores, the resulting scores correlated more strongly with the group judgments than with the individual judgments. Only for Judge 3 did the correlations of the individual judgments with the scores approach that of the group judgments. The two ways of using the ratings produced essentially the same correlations with the group judgments, although the scores based on total rating points tended to have higher correlations with the individual judgments than did the scores based on number of criteria passed. Unequal weighting of the criteria actually produced slightly lower correlations with the group judgments than did equal weighting.

After reviewing the analyses shown in Tables 4 and 5, the Praxis III developers decided to compute accuracy scores without penalty for differences of 0.5. For each of the other decisions involving the method of scoring, they chose the simpler method of computing the scores, basing the documentation score on the sum of the rating points and using equally weighted criteria for both the accuracy scores and the documentation scores. The remaining analyses in this report will focus exclusively on those accuracy and documentation scores.

Describing the relationship between scores and judgments

For setting a standard, the key question to be answered from the data produced by a study such as this one is, "What is the relationship between the numerical score assigned to a performance (in this case, a completed Record of Evidence form) and the probability that the performance will be judged acceptable?" One statistical procedure commonly used to estimate this kind of relationship is called logistic regression. This procedure assumes that the relationship can be described on a graph by a curve of a particular shape -- the shape of the curves in Figures 1 and 2. The data determine the extent to which this curve is shifted left or right and the extent to which it is compressed or elongated.[4] Although this assumption may not be literally true, the logistic regression technique typically provides good estimates of the statistical relationship in the region where the probability is neither extremely low nor extremely high.

---

[4]This curve has the mathematical equation

$$P = \frac{1}{1 + e^{-(a+bx)}},$$

where P is the probability, x is the score, e is the mathematical constant 2.71828..., and a and b are parameters estimated from the data.

When this type of analysis is used for setting a standard, the individuals responsible for choosing the standard often focus on the score for which the probability of a favorable judgment is .50. The reason for this choice is that above this score, the majority of the judgments tend to be favorable; below this score the majority of the judgments tend to be unfavorable. This score, which could be described as the "indifference point", will be referred to in this report as "$X_{.50}$".

Figure 1 shows the logistic regression curves that describe the estimated relationship between the accuracy scores and the accuracy judgments of the five individual judges. These curves reflect the substantial differences between the individual judges in their tendency to be strict or lenient. The curve for Judge 4, who classified only four of the fifteen Record of Evidence forms as acceptably accurate, is farthest to the right on the graph. Only a Record of Evidence form with a very high accuracy score would have a good chance of being accepted by this judge. The curve for Judge 3, who classified thirteen of the fifteen forms as acceptably accurate, is farthest to the left on the graph.

There were also differences between judges in the extent to which their judgments clearly implied a particular score as the standard for the accuracy scores. The slope of the curve for Judge 5 is quite steep, clearly implying a standard between 85 and 90. The slope of the curve for Judge 2 is much less steep, implying a standard somewhere between 75 and 95.

Figure 1 also implies that the data for Judge 4 (and, to a lesser extent, Judge 2) cannot be described well by a logistic regression curve. Notice that the curve for Judge 4 does not approach 1.00 as the accuracy score approaches 100; the logistic regression analysis does not take into account the fact that a score of 100 represents perfect accuracy. However, the inaccuracies that result from these limitations of the analysis are small in relation to the differences between the standards of the individual judges.

Table 6 shows the estimated probability of a favorable judgment from each of the five judges, for accuracy scores of 60, 65, etc. A favorable judgment, in this case, was a judgment that the criterion scores on the Record of Evidence form were acceptably close to the juried criterion scores. The estimated probabilities reflect the large differences between the individual judges. For example, a Record of Evidence form with an accuracy score of 80 would have a 92 percent probability of a favorable judgment from Judge 3 but only a 6 percent probability of a favorable judgment from Judge 4 and almost no chance of a favorable judgment from Judge 5. Table 6 also shows the "$X_{.50}$" score computed from the judgments of each judge. The judges' $X_{.50}$ standards range from about 73.5 to 92.0, with a mean of 83.6.

Figure 2 shows the logistic regression curve that describes the relationship between the documentation score and the group consensus documentation judgments. A favorable judgment, in this case, was a judgment that the documentation on the Record of Evidence form reflected acceptable performance as an assessor. Table 7 shows the estimated probability of a favorable consensus judgment from the group, for documentation scores of 50, 55, etc. The table also shows the estimated "$X_{.50}$" score of approximately 70.

Implications of the results

Choosing a passing score on a test is a policy decision. A standard-setting study provides one kind of information for making that decision; it identifies the the passing score implied by the personal standards of the individuals who serve as judges in the study. Other kinds of information may also enter into the decision, and these other kinds of information may imply a passing score different from that implied by the results of the study.

For the documentation scores, the results of the present study have some clear implications for the choice of a passing score. The judges were able to reach consensus for most of the Record of Evidence forms to be judged, and the consensus judgments implied a passing score of approximately 70. A documentation score of 65 implied a probability of less than .20 for a favorable judgment from the group of judges in the study; a documentation score of 75 implied a probability greater than .80. The range of passing scores that can reasonably be regarded as consistent with the results of the study is quite narrow.

For the accuracy scores, the results of the study did not clearly imply a passing score. The accuracy judgments of the most lenient judge implied a passing score in the range of 70 to 75; those of the severest judge implied a passing score in the range of 90 to 95. Therefore, the range of passing scores that could be interpreted as consistent with the results of the study is quite large. The limited availability of the judges did not allow time for them attempt to reach consensus in their accuracy judgments -- but it is not clear that any amount of time would have been sufficient.

Subsequent forms of the Assessor Proficiency Test will use the same scoring rules and formulas as the form used in this study. However, because each form will be based on a different videotaped performance, the forms may vary in difficulty. If they do, the accuracy and documentation scores required for certification should differ from one form to another. Generally, the best way to determine comparable standards on different forms of a test is to perform a statistical equating of the scores. However, the time required to take the Assessor Proficiency Test, the small numbers of assessor trainees trained at one time, and the lack of any other measure of the same proficiencies make it unlikely that an equating study will be possible. Without an equating study to provide a basis for choosing the required accuracy and documentation scores on subsequent forms, it may be necessary to replicate the standard-setting study with each new form of the test.

Table 1.
The Praxis III criteria.

Domain A - Organizing Content Knowledge for Student Learning

A1: Becoming familiar with relevant aspects of students' background knowledge and experiences

A2: Articulating clear learning goals for the lesson

A3: Demonstrating an understanding of the connections between the content that was learned previously, the current content, and the content that remains to be learned in the future

A4: Creating or selecting teaching methods, learning activities, and instructional materials or other resources that are appropriate for the students and that are aligned with the goals of the lesson

A5: Creating or selecting evaluation strategies that are appropriate for the students and that are aligned with the goals of the lesson

Domain B - Creating an Environment for student Learning

B1: Creating a climate that promotes fairness

B2: Establishing and maintaining rapport with students

B3: Communicating challenging learning expectations to each student

B4: Establishing and maintaining consistent standards of classroom behavior

B5: Making the physical environment as safe and conducive to learning as possible

Domain C - Teaching for Student Learning

C1: Making learning goals and instructional procedures clear to students

C2: Making content comprehensible to students

C3: Encouraging students to extend their thinking

C4: Monitoring students' understanding of content through a variety of means, providing feedback to students to assist learning, and adjusting learning activities as the situation demands

C5: Using instructional time effectively

Domain D - Teacher Professionalism

D1: Reflecting on the extent to which the learning goals were met

D2: Demonstrating a sense of efficacy

Table 1 (continued).
The Praxis III criteria.

D3: Building professional relationships with colleagues to share teaching insights and coordinate learning activities for students

D4: Communicating with parents or guardians about student learning

Table 2.
The accuracy judgments.

Record ID	Judge ID 3	1	2	5	4	Number OK
9	0	0	0	0	0	0
13	1	0	0	0	0	1
12	1	0	0	0	0	1
14	1	0	0	0	0	1
15	1	0	1	0	0	2
7	0	1	1	0	0	2
1	1	1	0	1	0	3
4	1	1	1	0	0	3
11	1	1	0	0	1	3
2	1	1	1	1	0	4
5	1	1	0	1	1	4
6	1	1	1	1	0	4
10	1	1	1	1	0	4
8	1	1	1	1	1	5
3	1	1	1	1	1	5
Number OK	13	10	8	7	4	

Correlations:	Judge 3	Judge 1	Judge 2	Judge 5	Judge 4
Judge 3	1.00	.14	.03	.37	.24
Judge 1	.14	1.00	.47	.66	.43
Judge 2	.03	.47	1.00	.34	-.04
Judge 5	.37	.66	.34	1.00	.34
Judge 4	.24	.43	.-.04	.34	1.00

Table 3.
The documentation judgments.

Record ID	Judge ID 1	3	5	2	4	Number OK	Group consensus[5]
13	0	0	0	0	0	0	0
12	0	1	0	0	0	1	0
14	0	1	0	0	0	1	1
15	0	0	0	1	0	1	0*
9	1	1	0	0	0	2	0
1	0	0	1	0	1	2	0
2	1	0	1	0	0	2	0
7	1	0	0	1	1	3	0*
4	1	1	0	1	0	3	0
11	1	1	0	0	1	3	1
5	1	0	1	0	1	3	1**
8	1	1	1	0	0	3	1
6	1	1	1	1	0	4	1
10	1	1	1	1	1	5	1
3	1	1	1	1	1	5	1
Number OK	10	9	7	6	6		

Correlations:

	Before discussion Judge 1	Judge 3	Judge 5	Judge 2	Judge 4	Group
Judge 1	1.00	.29	.38	.29	.29	.38
Judge 3	.29	1.00	-.05	.11	-.17	.50
Judge 5	.38	-.05	1.00	.05	.33	.46
Judge 2	.29	.11	.05	1.00	.17	.05
Judge 4	.29	-.17	.33	.17	1.00	.33
Group	.38	.50	.46	.05	.33	1.00

* Judge 2 disagreed with this group judgment.
** Judge 3 disagreed with this group judgment.

[5]Excluding Judge 5, who was not able to participate in the group discussion.

Table 4.
Comparison of scoring methods: Accuracy Scores

Penalty for difference of 0.5	0.5	0.5	None	None
Criterion weights	Equal	Unequal	Equal	Unequal
Highest score	91	91	100	100
Lowest score	58	58	67	70
Mean	72.9	73.9	85.2	86.5
Standard Deviation	8.9	8.3	9.3	8.4
Correlation with judgments				
Judge 1	.67	.62	.71	.68
Judge 2	.54	.56	.53	.58
Judge 3	.53	.44	.61	.53
Judge 4	.55	.50	.51	.45
Judge 5	.76	.74	.71	.67
Mean of correlations	.60	.56	.62	.60

Table 5.
Comparison of scoring methods: Documentation Scores

Basis for scores	Total Points	Total Points	Criteria Passed	Criteria Passed
Criterion weights	Equal	Unequal	Equal	Unequal
Highest score	83	85	95	97
Lowest score	54	53	47	45
Mean	68.6	69.2	73.7	74.3
Standard Deviation	9.1	9.7	16.5	16.9
Correlation with judgments Before discussion:				
Judge 1	.15	.10	.05	.00
Judge 2	.34	.31	.27	.26
Judge 3	.69	.66	.71	.49
Judge 4	.23	.22	.09	.05
Judge 5	.23	.21	.13	.09
After discussion:				
Group	.74	.72	.75	.72

Table 6.
Probability of a favorable accuracy judgment.

Score	Judge 1	Judge 2	Judge 3	Judge 4	Judge 5
100	.99+	.92	.99+	.82	.99+
95	.99	.84	.99+	.61	.99+
90	.94	.71	.99+	.34	.93
85	.78	.53	.98	.15	.15
80	.42	.34	.92	.06	.01-
75	.13	.19	.69	.02	.01-
70	.03	.10	.31	.01	.01-
65	.01	.05	.08	.01-	.01-
60	.01-	.02	.02	.01-	.01-
$X_{.50}$ =	81.08	84.32	73.49	92.00	87.02

Mean $X_{.50}$ = 83.58

Table 7.
Probability of a favorable documentation judgment.

Score	Probability
95	.99+
90	.99+
85	.99
80	.97
75	.86
70	.53
65	.18
60	.40
55	.01
50	.01-

$X_{.50} = 69.59$

Figure 1.

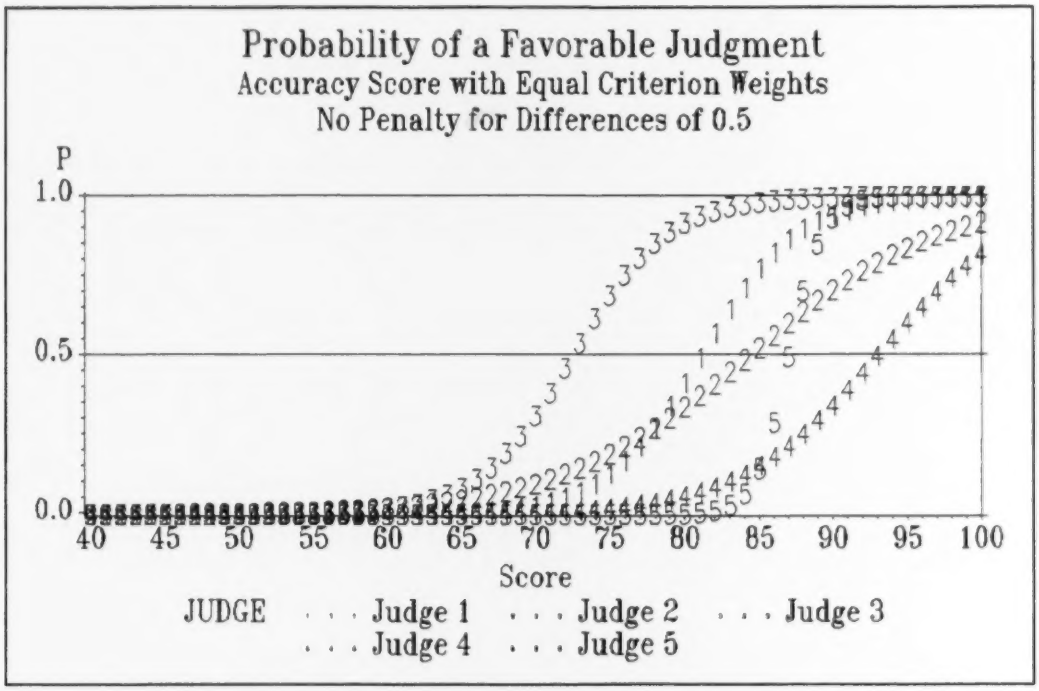
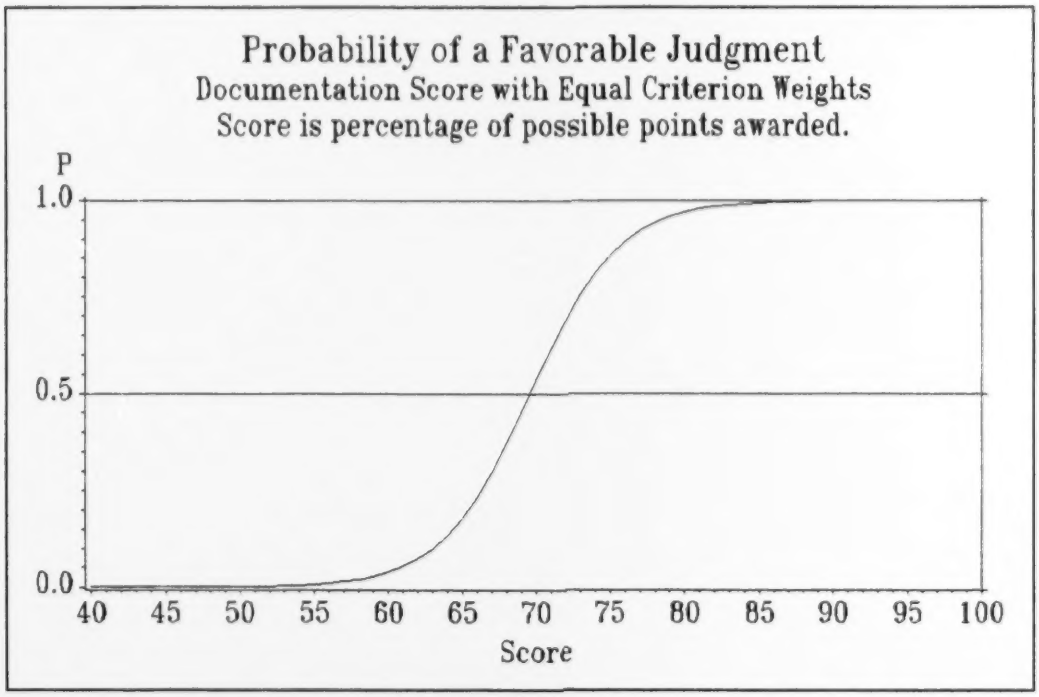

Probability of a Favorable Judgment
Accuracy Score with Equal Criterion Weights
No Penalty for Differences of 0.5
P
1.0
0.5
0.0
40 45 50 55 60 65 70 75 80 85 90 95 100
Score
JUDGE . . . Judge 1 . . . Judge 2 . . . Judge 3
. . . Judge 4 . . . Judge 5

Figure 2.


Probability of a Favorable Judgment
Documentation Score with Equal Criterion Weights
Score is percentage of possible points awarded.
P
1.0
0.5
0.0
40 45 50 55 60 65 70 75 80 85 90 95 100
Score